Using Excel to generate empirical sampling distributions

Rodney Carr
Deakin University, Faculty of Business and Law
PO Box 432,
Warrnambool, 3280, Australia
rodneyc@deakin.edu.au

Scott Salzman
Deakin University, Faculty of Business and Law
PO Box 432,
Warrnambool, 3280, Australia
scotts@deakin.edu.au

1. Introduction

Teachers in many introductory statistics courses demonstrate the Central Limit Theorem by using a computer to draw a large number of random samples of size n from a population distribution and plot the resulting empirical sampling distribution of the sample mean. There are many computer applications that can be used for this (see, for example, the Rice Virtual Lab in Statistics: http://www.ruf.rice.edu/~lane/rvls.html). The effectiveness of such demonstrations has been questioned (see delMas et al (1999))) but in the work presented in this paper we do not rely on sampling distributions to convey or teach statistical concepts; only that the sampling distribution is independent of the distribution of the population, provided the sample size is sufficiently large.

We describe a lesson that starts out with a demonstration of the CTL, but sample from a (finite) population where actual census data is provided; doing this may help students more easily relate to the concepts – they can see the original data as a column of numbers and if the samples are shown they can also see random samples being taken. We continue with this theme of sampling from census data to teach the basic ideas of inference. We end up with standard resampling/bootstrap procedures.

We also demonstrate how Excel can provide a tool for developing a learning objects to support the program; a workbook called Sampling.xls is available from www.deakin.edu.au/~rodneyc/PS > Sampling.xls.

W

2. Step 1 – Showing that the sampling distribution of the mean is the same for different populations

Demonstration of the Central Limit Theorem typically involve the sampling distribution of the sample mean with students being able to select different population distributions and see that the same shape is obtained no matter what the population distribution looks like, provided the sample size is sufficiently large. We do the same in this initial step, but in addition we make it clearer to students that the sampling distribution is the same for populations with the same mean and standard deviation. (It is not, for this exercise, important to examine the properties of the sampling distribution, such as showing that the standard deviation of the sampling distribution of the sample means is $\sigma/\sqrt{n}$.) We do this by providing a number of different populations all with the same μ and σ. The images in figure 1 show empirical sampling distributions (based on 5000 sample of size 100) for sample means drawn from 4 populations in Sampling.xls.

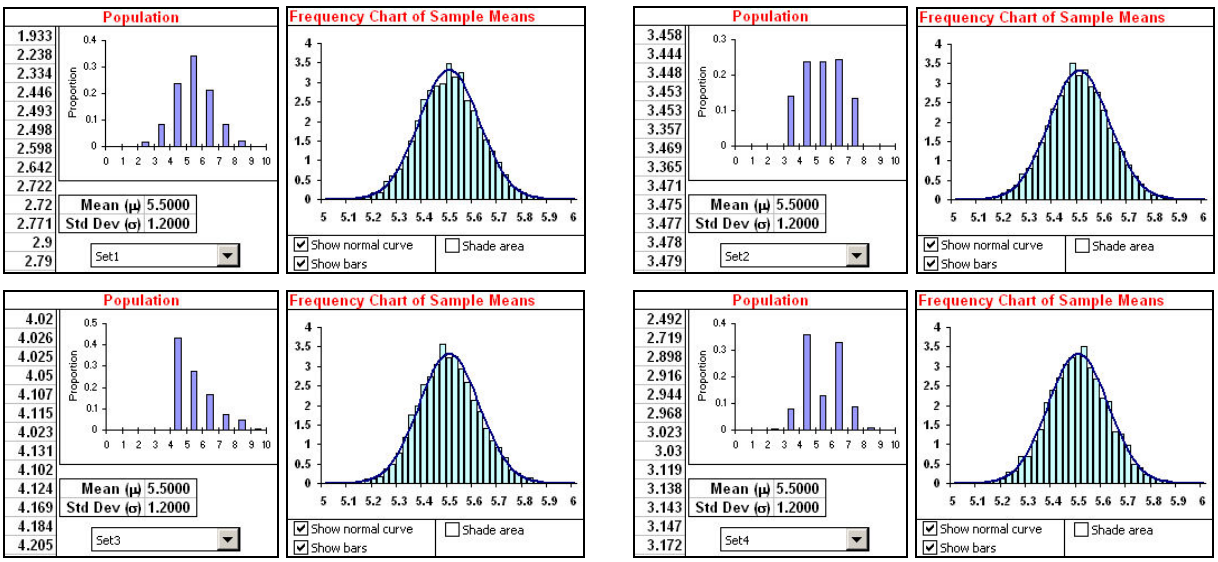


Figure 1. Sample distributions from populations with the same μ and σ

3. Step 2 – Making inferences using an empirical sampling distribution

This is standard inference, but without using the theoretical sampling distribution. The lesson may involve a series of questions a)-d) as follows:

a) What is the probability of drawing a sample of size 100 with a mean of over 5.6 from any of the populations in step 1?

The answer is the area under the histogram beyond 5.6 – the shaded area in figure 2.

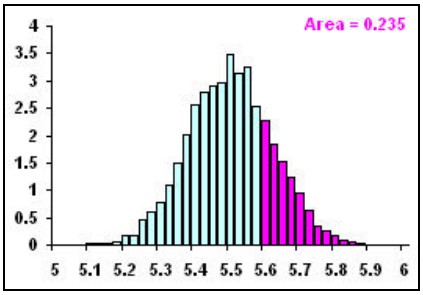


Figure 2. Probability of drawing a sample with a mean of over 5.6

b) What is the probability of getting a sample with a mean of over 5.9 from any of the populations?

Students have no problem answering zero.

The next question leads directly to the ideas of inference:

c) What would you conclude if you did draw a random sample with a mean of 5.9?

The answer is, of course, "It didn't come from any population with a mean of 5.5 and a standard deviation of 1.2.". Thus, we have the students drawing inferences based on a sample.

4. Step 3 – Showing the effect of different standard deviation in the population

Instead, the sample in question c) above might have come from a population with a different mean or standard deviation, such as the one shown in Figure 3 (that has a larger standard deviation).

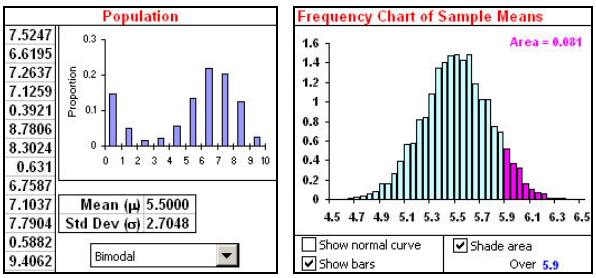


Figure 3. Probability of drawing a sample with a mean of over 5.6

5. Step 4 – Statistical inference using a simulated population

Now we ask a question such as:

d) The following sample is drawn randomly from a population.

3.2, 0.9, 2.8, 2.1, 1.6, 3.1, 3.5, 2.7, 2.5, 2.3, 2.4, 2.1, 0.8, 1.9,
2.2, 1.5, 2.8, 1.5, 2.5, 3.3, 1.8, 2.4, 0.8, 4.1, 2.5, 2.8, 2.5, 1.3,
1.8, 2.8, 2.3, 0.7, 2.6, 0.9, 2.6, 2.4, 2.8, 2.8, 3.4, 3.1

The sample mean is $\bar{x}$ = 2.3 and the sample standard deviation is s = 0.8. Can we conclude that the mean of the population is greater than 2.0? That, carry out the following hypothesis test:

H_0: $\mu = 2.0$

H_1: $\mu > 2.0$.

Since, after Step 1, students should now be convinced that, within reason, it doesn't matter what shape the population is, a natural way of answering the question is to construct an empirical sampling distribution by repeatedly drawing samples from a population with a mean of 2.0. Unfortunately, we don't know the population standard deviation, and this is clearly important (after Step 3). But if we assume that the sample is a 'representative' one, we can use the standard deviation of the sample, s, as an estimate of σ. Then we can simulate a population, just about any population, with a mean of 2.0 and a standard deviation of 0.8. Students can do this 'by hand' using the RAND and NORMINV functions in Excel, or use a tool such as SampSel in the XLStatistics collection (www.deakin.edu.au/~rodneyc/XLStats.xls). Figure 4 shows an image of an empirical sampling distribution drawn from such a simulated population.

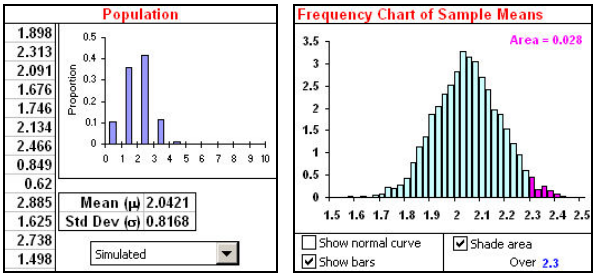


Figure 4. Sampling distribution using a simulated population

The answer to the question d) is, therefore "Since there is only about a 2.8% change of getting a sample like we have from a population with a mean of 2, and a standard deviation of 0.8, it is unlikely that the population has a mean of 2.0. That is, reject H_0 in favour of H_1". The students are now carrying out rigorous inferences, with no mention of theoretical distributions.

6. Step 5 – Resampling/bootstraping

The idea in this last step is that, instead of creating a simulated population using a tool like SampSel in XLStatistics, we can use the sample itself. The assumption is that the distribution of the population is similar to the distribution in the sample. Assuming this is not a lot stronger than assuming that the population standard deviation is approximately the same as the sample standard deviation. We do, however, need to adjust the sample to have the hypothesised value for the

population mean. Using the sample presented in the previous step, we subtract 0.3 (= 2.3 - 2.0) from each of the sample values and bootstrap from the resulting set. Figure 5 shows the resulting sample distribution.

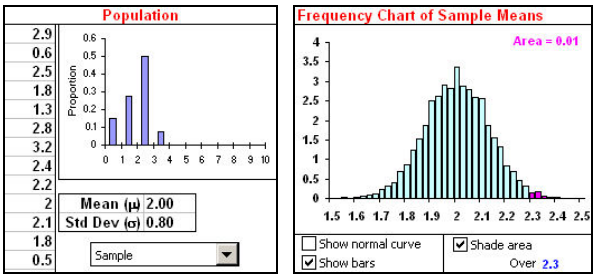
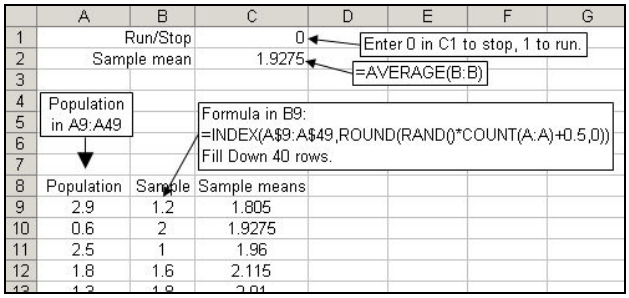


Figure 5. Bootstrapping to obtain a sampling distribution

The conclusion is the same as that made in Step 4.

7. Doing it using Excel

All the above images are from Excel. The essential information needed to draw a random sample with replacement of size n is shown in figure 6.




```
Sub SetOnCalculate()
    ActiveSheet.OnCalculate = "ReCalculate"
End Sub
Sub ReCalculate()
    While Range("C1") <> 0
        Range("C" & Range("C1") + 8) = Range("C2")
        Range("C1") = Range("C1") + 1
        DoEvents
    Wend
End Sub
```

Figure 6. Taking a random sample of size n = 40 without replacement

The VB code pictured in figure 6 is entered into a new Module. Run 'SetOnCalculate'. Then altering the value in C1 to 1 will cause samples to be taken one after another. The sample means are recorded in Column C and a frequency chart can be drawn to show the sampling distribution of the sample means.

If necessary samples can be generated without replacement – see, for example Christie (2004).

REFERENCES

delMas RC, Garfield J, and Chance BL (1999), A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning , *Journal of Statistics Education* v.7, n.3

Derek Christie, Resampling with Excel, *Teaching Statistics*, 26(1), p9-14.